Walmart Weekly Sales Forecast on Store-Department Level

...

Xiaohan Liu
@ OneCareer
Aug-22-2024

Roadmap

1. Data Preprocessing
    a. Tables integration
    b. Missing values handling
    c. Outliers handling
2. Feature Engineering
    a. Replace Markdown 1~5 by their summation
    b. Replace IsHoliday by IsSB, IsLD, IsTG, IsXM to discriminate holidays
    c. Replace week_of_year to Wks2TG and Wks2XM
    d. Add preceding two weeks' sales as lag_1 and lag_2
3. Modeling
    a. LightGBM
    b. SARIMAX
    c. Prophet
4. Evaluation and Discussion

Data Preprocessing

Table Integration

With the help of external sourced lexicon of the department codes, we could combine some departments by their descriptions.

Results:

1. Reduce the department count in our dataset from 81 to 66.

Department Name - the department name corresponding to the department number. Some department codes share the SAME department name, but not necessarily the same In-Full Category and Merch Align. We find some of the department names are not unique, and some department names are not present in the original dataset. We will substitute the department names that are not present in the original dataset with the department numbers that have the same department names in the original dataset. After this department grouping, we will have 71 departments in the walmart_dept dataset.

- 'Stationery' is named as 03 Stationery, covering 3 departments [3, 53, 73].
- 'Piece Goods Fabrics and Crafts' is named as 19 Piece Goods Fabrics and Crafts, covering 3 departments [19, 44, 52].
- 'Jewelry' is named as 32 Jewelry, covering 3 departments [32, 47, 54].
- 'Sporting Goods' is named as 09 Sporting Goods, covering 3 departments [9, 45, 51].
- 'Cosmetics Fragrances and Skin Care' is named as 46 Cosmetics Fragrances and Skin Care, covering 2 departments [46, 59].
- 'Auto Service' is named as 37 Auto Service, covering 2 departments [37, 65].
- 'Mens Wear' is named as 23 Mens Wear, covering 2 departments [23, 41].
- 'Optical' is named as 49 Optical, covering 2 departments [49, 58].
- 'Ladies Wear' is named as 34 Ladies Wear, covering 2 departments [34, 40].
- 'Deli' is named as 80 Deli, covering 2 departments [80, 97].
- 'Automotive' is named as 10 Automotive, covering 2 departments [10, 42].
- 'Toys' is named as 7 Toys, covering 2 departments [7, 43].
- 'Media and Gaming' is named as 5 Media and Gaming, covering 2 departments [5, 55].

Table Integration

With the help of external sourced lexicon of the department codes, we could combine some departments by their descriptions.

Results:

2. Potential Store-Category/Merch_Align analysis.

	Department Name	Accounting Department	In Full Category	Merch Align
0	Candy and Tobacco	1	Food	Food
1	Personal Care Health and Beauty Aids	2	Consumables	Consumables
2	Stationery	3	General Merchandise	Hardlines
3	Paper Goods Household Paper	4	Consumables	Consumables
4	Media and Gaming	5	General Merchandise	Entertainment, Toys, and Seasonal
...	...	...	...	...
66	Fresh Produce	94	Food	Food
67	DSD Grocery and Snacks	95	Food	Food
68	Liquor Adult Beverage	96	Food	Food
69	Bakery	98	Food	Food
70	Office and Store Supplies	99	General Merchandise	Hardlines

Store	New_Dept	num_dept
1	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	65
2	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	66
3	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	60
4	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	66
5	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	60
...	...	...
41	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	65
42	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	56
43	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	56
44	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	57
45	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	62

Table Integration

- Given the fact that not all stores possess the same set of departments, we will first enumerate all store-department pairs for each store.

- Then, left join with the walmart_dept table to get the department name, In-Full Category, and Merch Align information for each store-department pair.

- It is ensued by left joining with the stores table to get the store type and size information for each store-department pair. Up to this step, we have a 2,821 by 7 dataframe, which covers all the possible store-department pairs and their corresponding department name, In-Full Category, Merch Align, store type, and size information.

- Next, we right join with the features table to get all the features at all weeks during the selected time period for each store-department pair, even though this span causes some missing values for 513,422 store-dept pairs' weekly sales. We will handle this during missing data imputation later.

- Eventually, we left join with the train table to get the weekly sales for each store-department pair. Remember to drop the duplicate 'isHoliday' column from the train table, since it is already in the features table.

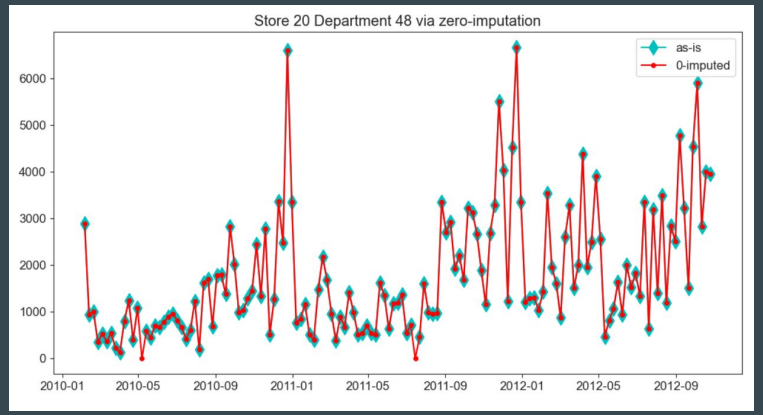
- As the final touch, we drop the weeks greater than the max date, 2012-10-26, in the train data, since we don't have sales data for those weeks. At last, the joined dataframe is a 403,403 entries by 19 columns table.

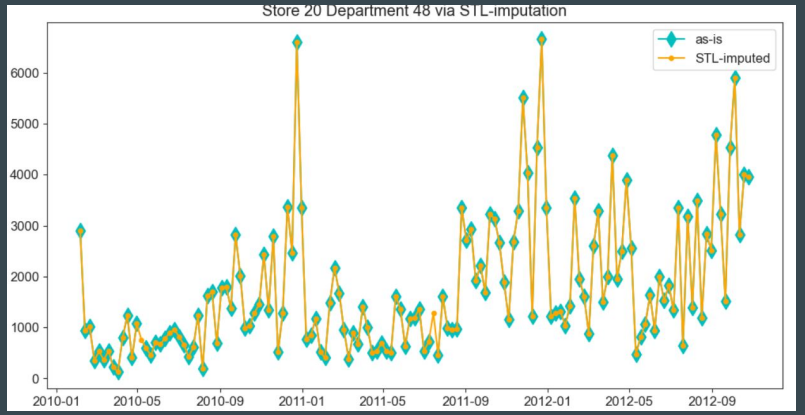
Conclusion: We decide to work on individual store-department pair level of analysis for now and grow to store level or department level analysis.

Missing Values

-Due to lack of ascertain on the missing mechanism, we use 0 to impute the missing values in markdowns.And use both 0-imputation and STL-imputation on the missing Weekly_Sales for eligible store-department pairs.

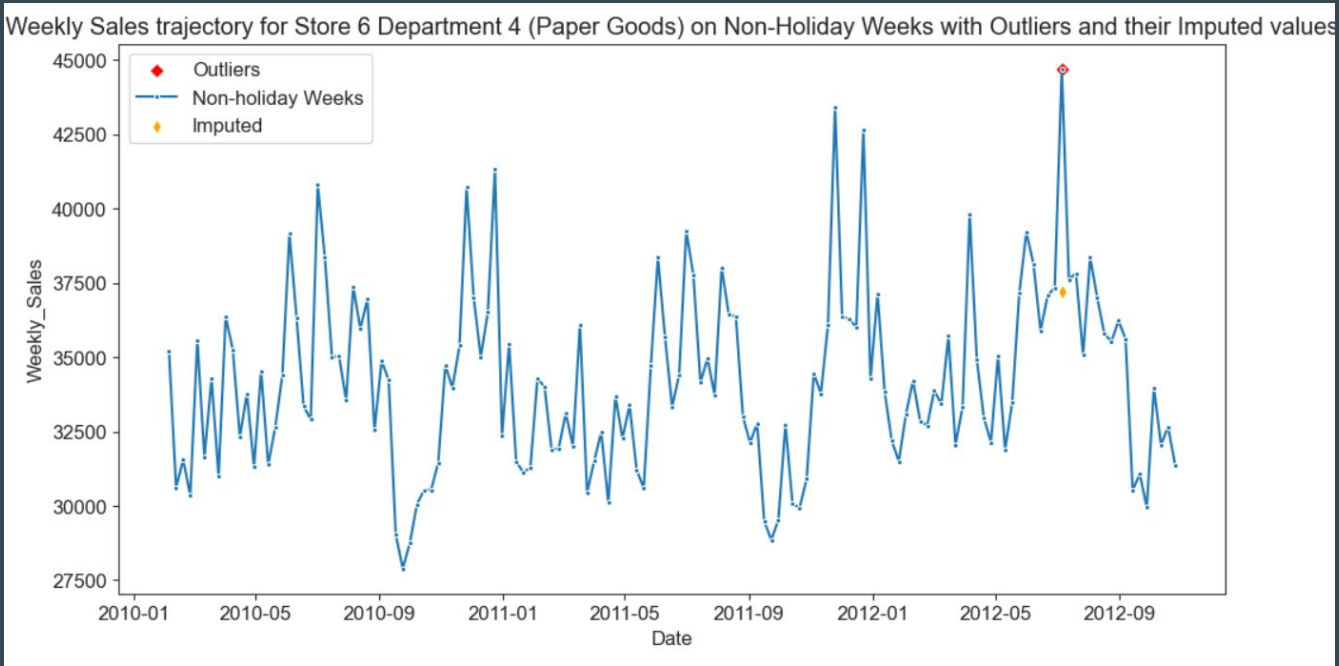
- Our analysis is conducted on 3 missing-value-free pairs: store-20(A)-department-4, store-13(A)-department-4 and store-6(A)-department-4. Analysis on other pairs could be done ad hot whenever needed.





Outliers

- Exclude the holidays and 2 weeks before holidays from the screening.
- Use Tucky's fence to detect and rolling mean from previous 3 to 4 weeks to impute outliers.



Feature Engineering

Feature Engineering

a. Replace Markdown 1~5 by their weekly summations

b. Replace column IsHoliday by boolean columns IsSB, IsLD, IsTG, IsXM to discriminate among holidays

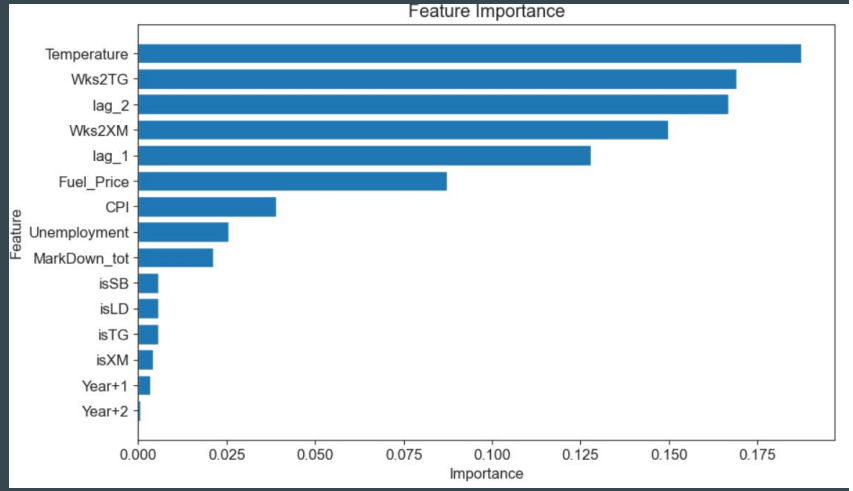
c. Replace week_of_year to Wks2TG and Wks2XM, so we have 2 numeric features instead of a categorical one with high cardinality.

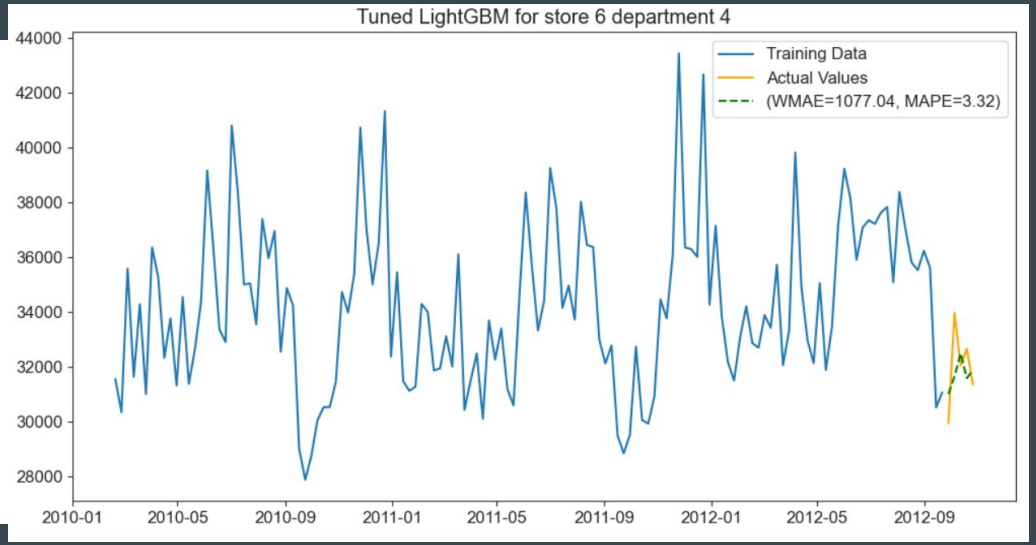
d. Add preceding two weeks' sales as lag_1 and lag_2 (Perhaps more is preferred for non-time series specific models).

e. Normalize the numeric features.

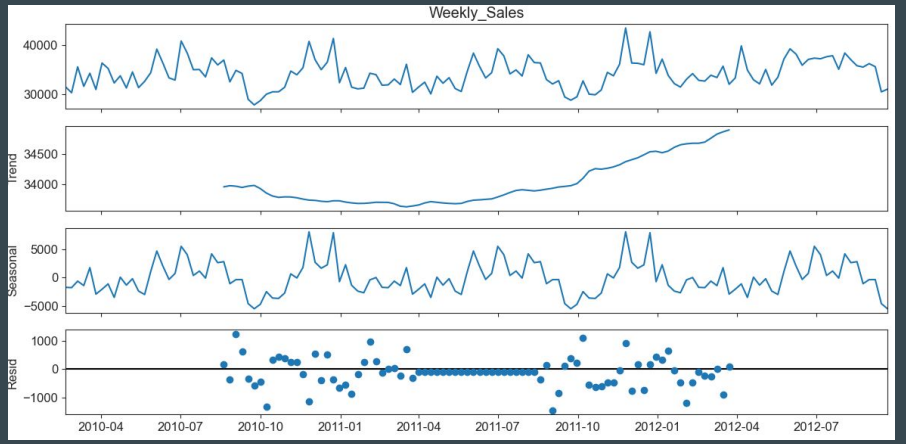
Modeling

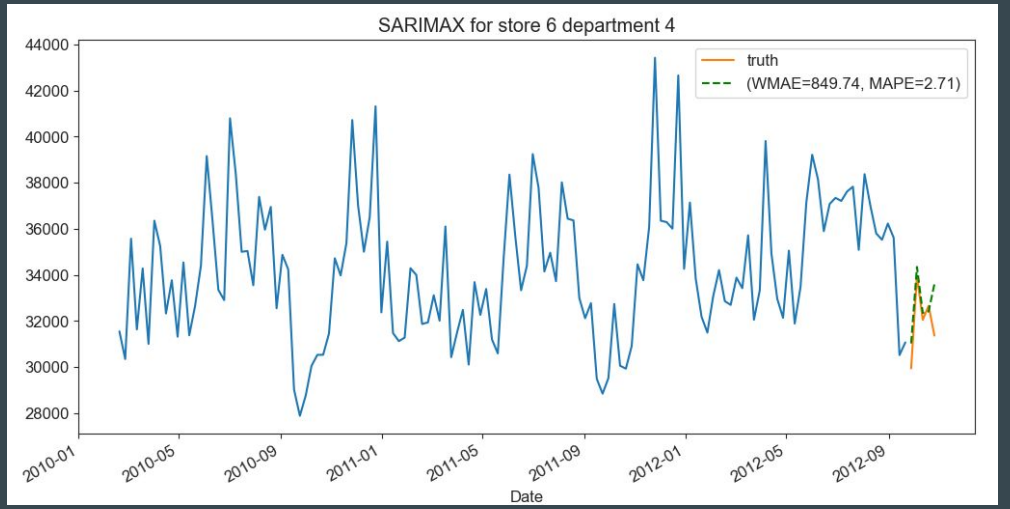
LightGBM on Store 6 Department 4



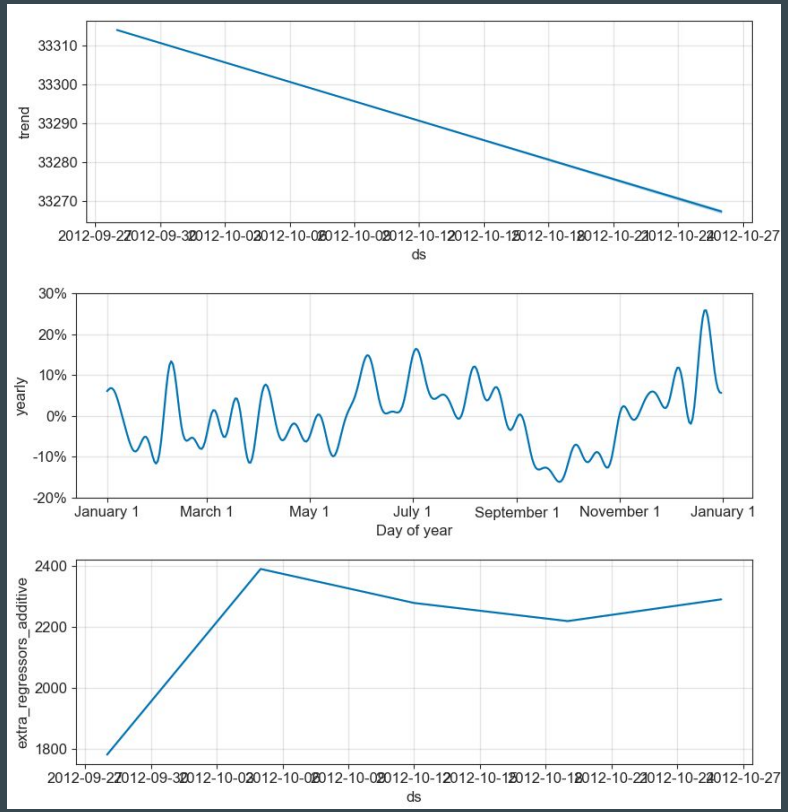


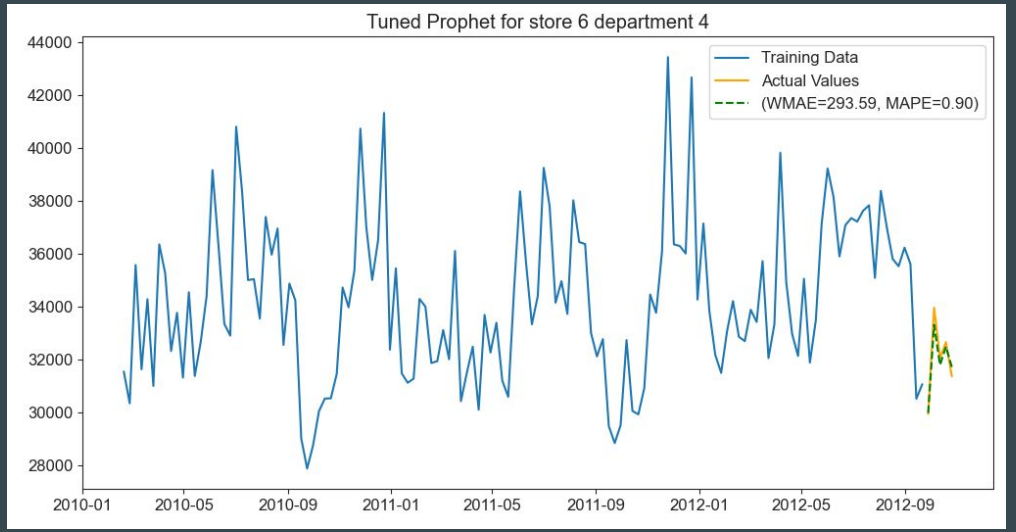
SARIMAX on Store 6 Department 4





Prophet on Store 6 Department 4





Evaluation & Discussion

- Time-series specific models like SARIMA and Prophet perform better than more generalized models like LightGBM.
- Patterns for each Store-Department might be different.
- Hyperparameter tuning is performed by random grid searches with cross-validation.
- Future works include implementing LSTM, more store-department level and/or store-category/merch_align analysis.

```
print(table_s6d4)
```

	Model	WMAE	MAPE
0	SARIMA	849.74	2.71
1	LightGBM	1077.04	3.32
2	Prophet	293.59	0.90

```
print(table_s13d4)
```

	Model	WMAE	MAPE
0	SARIMA	941.08	2.13
1	LightGBM	1218.03	2.68
2	Prophet	503.89	1.14

```
print(table_s20d4)
```

	Model	WMAE	MAPE
0	SARIMA	2700.10	4.95
1	LightGBM	2870.03	5.48
2	Prophet	2243.15	4.34

Thanks!